Tracing the Origin of Food-borne Disease Outbreaks

Jacobs, Rianne; Teunis, Peter; Van De Kassteele, Jan

Published in:
Epidemiology

DOI:
10.1097/EDE.0000000000001169

OPEN

Tracing the Origin of Food-borne Disease Outbreaks
A Network Model Approach

Rianne Jacobs,[a,b] *Peter Teunis,*[c] *and Jan van de Kassteele*[d]

Background: Food-borne disease outbreaks constitute a large health burden on society. One of the challenges when investigating such outbreaks is to trace the origin of the outbreak. In this study, we consider a network model to determine the spatial origin of the contaminated food product that caused the outbreak.

Methods: The network model we use replaces the classic geographic distance of a network by an effective distance so that two nodes connected by a long-range link may be more strongly connected than their geographic distance would suggest. Furthermore, the effective distance transforms complex spatial patterns into regular topological patterns, creating a means for easier identification of the origin of the spreading phenomenon. Because detailed information on food distribution is generally not available, the model uses the gravity model from economics: the flow of goods from one node to another increases with population size and decreases with the geographical distance between them.

Results: This effective distance network approach has been shown to perform well in a large *Escherichia coli* O104:H4 outbreak in Germany in 2011. In this article, we apply the same method to various food-borne disease outbreaks in the Netherlands. We found the effective distance network approach to fail in certain scenarios.

Conclusions: Great care should be taken as to whether the underlying network model correctly captures the spreading mechanism of the outbreak in terms of spatial scale and single or multiple source outbreak.

Keywords: Effective distance; Food-borne diseases; Network model; Source tracing

(*Epidemiology* 2020;31: 327–333)

Food-borne disease outbreaks constitute a large health burden on society.[1] The only way to mitigate this burden is to quickly and efficiently identify food-borne disease outbreaks and the contaminated food product that caused it.

The identification of a contaminated food product is a long and cumbersome process involving many steps. Case-control studies may be used to analyze the eating patterns of cases to identify the contaminated food product.[2] The case-control studies are traditionally analyzed using a combination of univariable and multivariable models and stepwise variable selection procedures.[3,4] A recent study has made a first attempt at formalizing the methodology needed in this identification process.[5]

Food-borne disease outbreaks often show a complex spreading pattern over large geographical areas. To find the cause of the outbreak, authorities need to know the food distribution network to trace the origin of the outbreak. However, it is often not possible to monitor specific food transportation routes and detailed food distribution information is generally not.[6]

One way to deal with these complications is to make use of a network model to reconstruct the outbreak origin. One such model was developed by Manitz et al.[6] The two main characteristics of this model are (1) replacing the geographic distance by an effective distance measure and (2) using the gravity model from.[7,8] The effective distance deals with the problem of the spatially complex spreading pattern of the outbreak, and the gravity model allows the reconstruction of a transportation network in the absence of food distribution data. Manitz et al[6] showed that their method worked well in reconstructing the outbreak origin of the *Escherichia coli* O104:H4 outbreak in Germany in 2011.[9] It is not clear, however, if this method is able to accurately reconstruct the origin for any outbreak, irrespective of its scale and spatial spreading pattern.

In this article, we attempt to answer this question by applying the network model to three food-borne disease

Submitted June 28, 2018; accepted January 24, 2020.

From the [a]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, the Netherlands; [b]Department of Statistics, Informatics and Modelling, National Institute for Public Health and the Environment, Bilthoven, Netherlands; [c]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA; and [d]Department of Statistics, Informatics and Modelling, National Institute for Public Health and the Environment, Bilthoven, Netherlands.

This work is partially supported by the National Institute for Public Health and the Environment (RIVM) and through their Strategic Programme (SPR), which contributes to solutions to societal challenges through interdisciplinary research and by supporting innovation and capacity building at RIVM.

The authors report no conflict of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Data and code can be obtained from the authors.

Correspondence: Rianne Jacobs, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, PO Box 407, 9700 AK Groningen, the Netherlands. E-mail: rianne.jacobs@rug.nl.

ISSN: 1044-3983/20/3103-0327
DOI: 10.1097/EDE.0000000000001169

outbreaks in the Netherlands, namely a *Salmonella* thompson outbreak in 2012,[4] *Salmonella* typhimurium outbreak in 2006,[10] and *Escherichia coli* O157 outbreak in 2007.[11] Each outbreak has been shown to have its own point source location in space. The 2012 and 2007 outbreaks showed a uniform pattern across the whole country, whereas the 2006 outbreak showed a clustered pattern in a specific region.

In the Method section, we present the network model. The Data section presents the three datasets and the Results section the results of the analyses. In the Discussion section, we discuss the results and provide the conclusions in the Conclusion section.

METHOD

To model spatial food distribution, we use a network model which consists of a set of nodes $\mathcal{K} = 1,2,\ldots,K$ representing administrative regions in the Netherlands, such as municipalities or neighborhoods. These nodes are connected by a set of links $\mathcal{L} = \{(k,l) \mid k,l \in \mathcal{K}\}$. The basic idea is that, given some effective distance definition, the spreading pattern of an outbreak represents a concentric pattern from the outbreak origin, $k_0 \in \mathcal{K}$.[12]

The effective distance is defined as

$$d_{\text{eff}}(k,l) = \min_{\gamma \in \Gamma_{k,l}} (L(\gamma) - logP(\gamma)) \quad (1)$$

where $\Gamma_{k,l}$ is the set of all possible paths γ from node l to node k, $L(\gamma)$ is the length of path γ given by the number of links in the path along the nodes $k_0, k_1, \ldots, k_{L(\gamma)}$ and $P(\gamma)$ is the path probability given by the product of the transition probabilities $p_{k_i,k_{i-1}}$ of the corresponding links in the path γ.[12]

Because transportation network data is often not available to calculate the transition probabilities, Manitz et al[6] suggested making use of the gravity model.[7,8] The gravity model assumes that the amount of goods flowing from one region to another increases with population size and decreases with the geographic distance between regions:

$$F_{k,l} \propto \begin{cases} \dfrac{N_l^{\alpha} N_k^{\beta}}{(1 + d_{k,l}/d_0)^{\gamma}} & k \neq l \\ 0 & k = l \end{cases} \quad (2)$$

where N_k and N_l denote the population sizes of regions k and l, respectively, $d_{k,l}$ the distance between regions k and l. The non-negative values α, β, γ, and d_0 are parameters of the model. Following Manitz et al,[6] we choose these to be $\alpha = \beta = 0.5$, $\gamma = 2.6$ and d_0 the average linear extent (radius) of a region. Derivation of these parameter values is given by Manitz et al.[6]

The transition probability from node l to node k is then given by

$$p_{k,l} = \frac{f_{k,l}}{\sum_l f_{k,l}} \quad (3)$$

where $f_{k,l}$ is the flux obtained as $F_{k,l} / (\sum_{kl} F_{k,l})$ with $F_{k,l}$ denoting the flow of goods from node l to node k as given in Equation 2.

The gravity model results in a fully connected network in which every node is connected to every other node. This, however, is not realistic, as food transport networks are usually quite.[6] We take this sparsity into account by only retaining links that are significantly different from a random null model.[13] If, for each node, traffic is randomly distributed among the remaining K nodes, a null model would give $p^0_{k,l} \approx 1/K$. We, therefore, only keep those links with a flux fraction greater than $1/K$: $p_{k,l} > 1/K$. The resulting graph captures the typical multi-scale structure of transportation networks, namely strong short links and a few long-range links.

For a given node, the shortest-path tree can be calculated. This is the collection of shortest effective paths (from Equation 1) to all other nodes in the network. This shortest-path tree constitutes the most probable hierarchy that a spreading process will take through the network.[6] The network-based origin estimation approach relies on the assumption that in the effective distance topology, only from the perspective of the actual outbreak origin, the network pattern represents a regular concentric wavefront structure[6] (eFigure 8; http://links.lww.com/EDE/B644). In combination with the observed disease pattern, which usually consists of a subset of nodes with nonzero incidence, the outbreak origin can be reconstructed by minimizing the expected value of the effective distance $\mu(d_{\text{eff}}; k_0)$ from the origin k_0 to all other nodes $k \in \mathcal{K}$ in the network with a nonzero incidence.[12]

Due to low population sizes in certain regions, the observed incidence may be unrealistically low or high. To avoid these extremes, we introduce a novel modification to the above method by fitting a generalized linear mixed effect model to the observed incidence, $\frac{n_k}{pop_k}$, in the following way

$$n_k \sim \text{Poisson}(\mu_k)$$

$$log\left(\frac{\mu_k}{pop_k}\right) = \alpha + \theta_k$$

$$\theta_k \sim N(0, \sigma_\theta)$$

with n_k the number of cases in region k and pop_k the population in region k. In this way, the incidence of each region is shrunken towards an overall mean incidence using information from the other regions via the random intercepts, θ_k.[14] Note that the population numbers used in the Poisson model above are needed to calculate the incidence needed for the model. This should not be confused with the use of the population numbers in the network model (Equation 2). The population numbers in the network model are needed to describe the flow of goods between regions, as required by the gravity model.

The expected distance can be estimated by the mean effective distance:

$$\hat{\mu}(d_{\text{eff}};k_0) = \frac{1}{N_x}\sum_{k=1}^{K} x_k d_{\text{eff}}(k,k_0) \quad (4)$$

where $N_x = \sum_k x_k$ is the sum of the estimated incidences $x_k = \exp(\alpha + \theta_k)$, $k = 1,2,\ldots,K$. Weighting the mean by the incidence, x_k, is a more robust alternative for noisy data than using only the process.[12]

A second novel contribution is the introduction of a scaling approach to obtain more accurate origin estimation. This is done by applying our method to three outbreaks in the Netherlands and perform the analysis on three different spatial aggregation levels, namely, municipality, district, and neighborhood. Municipality level is the highest level of aggregation, neighborhood the lowest. For each outbreak, we start at municipality level for the whole country. We then zoom in to lower aggregation levels. We obtained the data on the areas and population sizes for each of the municipalities, districts, and neighborhoods from Statistics Netherlands (CBS, 2018); the data are publicly available on their website (https://opendata.cbs.nl).

FOOD-BORNE DISEASE OUTBREAK DATA

Figure 1 illustrates the disease pattern of the three outbreaks and the region of the outbreak origin.

1. *Salmonella* thompson (2012) (951 cases) (4): A national outbreak with cases spread out across the whole country. The outbreak origin was located near the geographic center of the Netherlands. The contaminated food product was found to be smoked Salmon.
2. *Salmonella* typhimurium (182 cases) (2006) (10): A regional outbreak with the majority of cases clustered in the East of the Netherlands. The outbreak origin was also located in the East of the Netherlands. The contaminated food product was found to be a locally produced cheese.
3. *Escherichia coli* O157 (2007) (41 cases) (11): An international outbreak with cases in Iceland and the Netherlands. The cases in the Netherlands were spread out across the whole country. The suspected outbreak origin (no microbiological evidence was found) was located in the West of the Netherlands. The suspected contaminated food product was prepackaged lettuce.

RESULTS

The results are shown as choropleth maps, in which the color scale indicates the value of the mean effective distance $\hat{\mu}(d_{\text{eff}};k_0)$. The scale goes from yellow (high mean values) to red (low mean values). As low mean values indicate higher concentricity, the likelihood of a region to be the actual outbreak origin increases as we move along the color scale from yellow to red. The region with the lowest mean, i.e., the estimated outbreak origin, is colored in red. The region containing the suspected outbreak origin is indicated by a black cross.

Figure 2 shows the results for *Salmonella* thompson. Here, we only applied the origin reconstruction method on the municipality level because fitting the model on a finer aggregation level had no added value. At first sight, the origin reconstruction method seems to perform very well. The reconstructed outbreak origin is close to the actual outbreak origin (approximately 20 km). During the analysis, however, we doubted whether the method indeed performs well. Maybe the method actually just

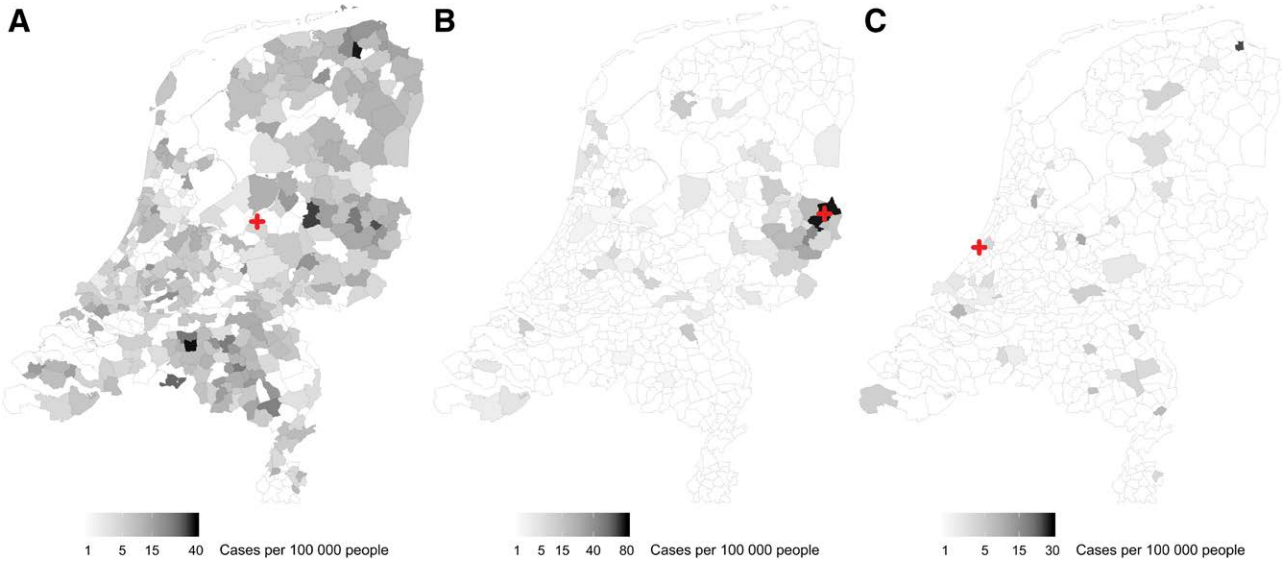

FIGURE 1. Disease patterns and outbreak origin of three food-borne disease outbreaks in the Netherlands. The municipality containing the outbreak origin is indicated by a red cross. The grayscale indicates the disease incidence with dark regions indicating high incidence.

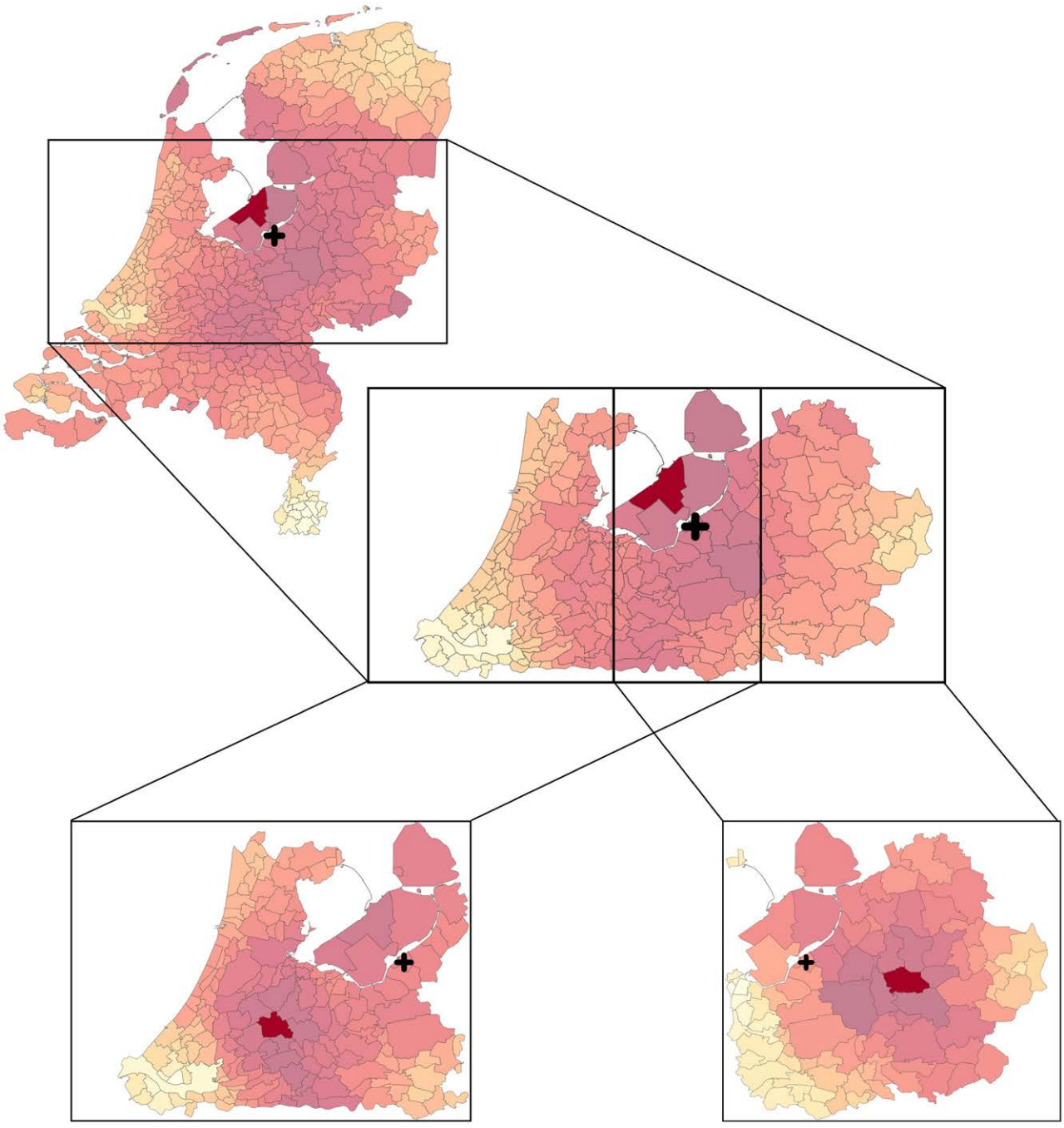

FIGURE 2. Results of reconstruction of outbreak origin for the *Salmonella* thompson outbreak data. The color gradient of the districts indicates the value of the weighted mean, with the red spectrum indicating low values and the yellow spectrum indicating high values. The black cross indicates the region containing the actual outbreak origin. The red district constitutes the reconstructed outbreak origin. The analysis was performed on municipality level. Each frame constitutes a subset of the data.

ends up in the center of gravity of the country, which just happens to be close to the actual outbreak origin. This latter point indeed seemed the case. When we consider a smaller geographical area of the country, such as the western and eastern parts, we note that the reconstructed outbreak origin again ends up in the middle of the analyzed area and not close to the actual origin outbreak.

Figure 3 shows the results for *Salmonella* typhimurium. We applied the origin reconstruction method on all three levels of the spatial hierarchy. For this outbreak, we see a completely different picture. Although the true source is located near the eastern border, the method is able to correctly identify the municipality where the outbreak originated. Moreover, when we

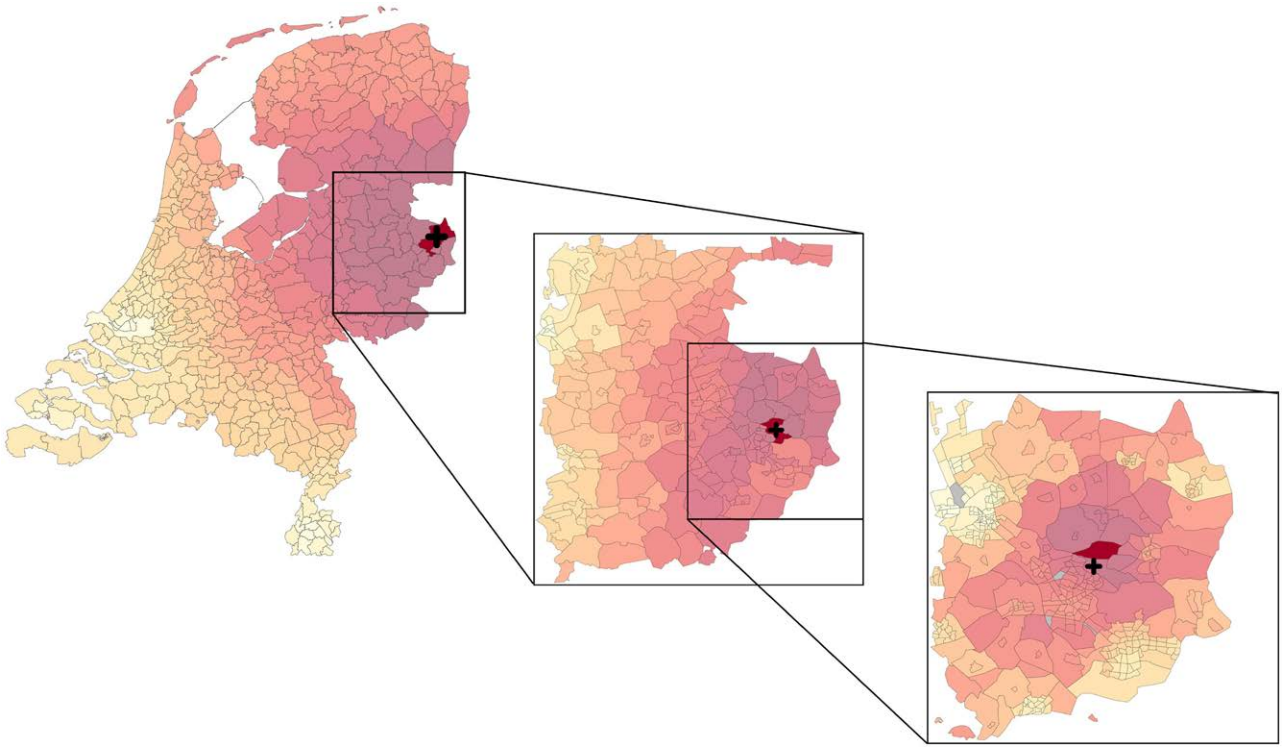

Figure 3. Results of reconstruction of outbreak origin for the *Salmonella* typhimurium outbreak data. The color gradient of the districts indicates the value of the weighted mean, with the red spectrum indicating low values and the yellow spectrum indicating high values. The black cross indicates the region containing the actual outbreak origin. The red district constitutes the reconstructed outbreak origin. From left to right, the data were analyzed on municipality, district, and neighborhood level.

zoom in and run the analysis on district level, the method correctly reconstructs the district where the outbreak originated. When we zoom in even more and run the analysis on neighborhood level, the method points to the neighborhood next to the neighborhood where the outbreak originated.

Figure 4 shows the results for *Escherichia coli*. Here, we only applied the origin reconstruction method on the municipality level. Our doubts about the performance of the method were confirmed when considering the *Escherichia coli* outbreak. Here, we see that the method reconstructs the outbreak origin to be in the middle of the country, although the most likely origin was very much on the west coast, approximately 90 km removed from the estimated origin. Again, when we consider a smaller geographical area, namely the western portion of the country, actually guiding the method to look more closely in the direction of the actual outbreak origin, the model still fails to correctly identify the actual outbreak origin.

For both the *Salmonella* thompson and the *Escherichia coli*, the falsely estimated areas lie in the center of the (subsection of) the map. The network, therefore, seems to favor the areas that are in the geographic center of the area under investigation.

DISCUSSION

From the above results and discussion, we deduce that the method to reconstruct the outbreak origin only works in certain cases. The most evident difference between the *Salmonella* thompson and *Escherichia coli* outbreaks, and the *Salmonella* typhimurium outbreak is the scale of the outbreak. Although all three were technically national outbreaks with confirmed cases found across the Netherlands, the *Salmonella* typhimurium outbreak clearly had a more clustered disease pattern. The disease pattern of the other two outbreaks was uniformly spread out across the Netherlands, and consequently, the spreading pattern did not contain sufficient information on the source of the outbreak. This was evidenced by the fact that the outbreak origin for these two outbreaks was estimated to be the same area, namely the geographic center of the Netherlands.

The above analyses only considered the final state of the outbreak, i.e., the disease pattern of all cases, ignoring the progression of the outbreak over time. To investigate the performance of the method when considering the spreading pattern of the outbreak, we also analyzed the data on a weekly basis (eFigures S.2–S.4; http://links.lww.com/EDE/B644). We found that the spreading pattern for the three outbreaks did not change over time, i.e., the *Salmonella* thompson and *Escherichia coli* outbreak had cases all over the country from

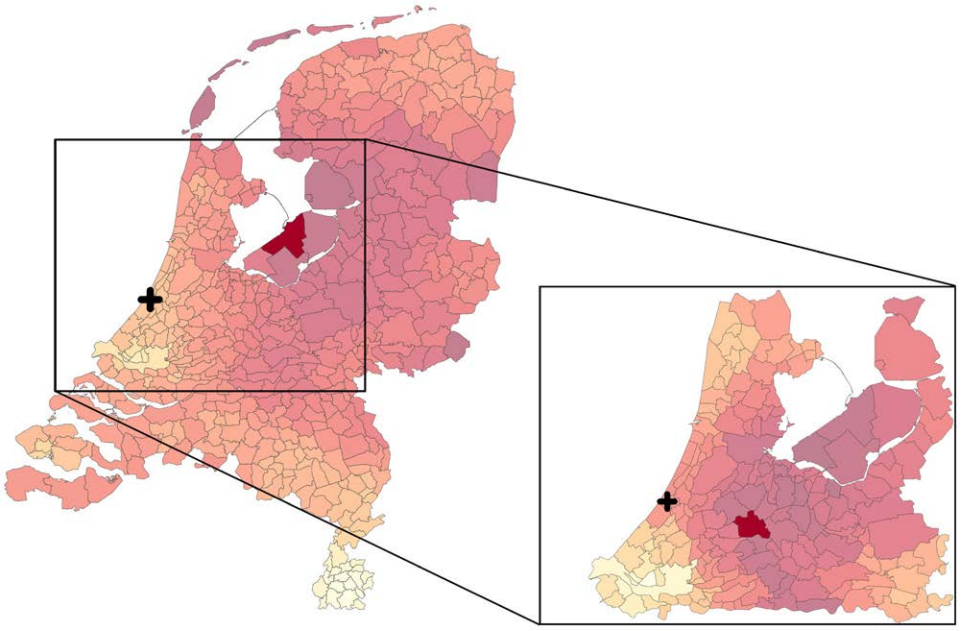

Figure 4. Results of reconstruction of outbreak origin for the *Escherichia coli* outbreak data. The color gradient of the districts indicates the value of the weighted mean, with the red spectrum indicating low values and the yellow spectrum indicating high values. The black cross indicates the region containing the actual outbreak origin. The red district constitutes the reconstructed outbreak origin. The analysis was performed on municipality level.

the start of the outbreak and the *Salmonella* typhimurium outbreak was clustered from the beginning. Therefore, even early in the outbreak, the method could not correctly estimate the outbreak origin for the *Salmonella* thompson and *Escherichia coli* outbreaks.

In the Method section, we mentioned our specific parameter choice for the parameters of the gravity model (Equation 2). Following the method of Manitz et al,[6] we chose these to be $\alpha = \beta = 0.5$, $\gamma = 2.6$ and d_0 the average linear extent (radius) of a region. The values $\alpha = \beta = 0.5$ are obtained by assuming that the coupling strength between two regions increases with the number of connections that can be formed and is proportional to the geometric mean.[6] We performed a sensitivity analysis for the parameter γ (eFigure S5; http://links.lww.com/EDE/B644). Our findings confirm those of Manitz et al,[6] namely that the results were quite robust against changes in this parameter. Ideally, one would like to estimate the parameters of the gravity model. We should also note that the specific parameter selection may not be generalizable to other scenarios.

The scaling of the network-based origin estimation is a good way to obtain a more specific estimate of the origin. This scaling is done by selecting a subset of regions in which the network model is then fitted. The model can be fitted using smaller regions such as neighborhoods instead of municipalities, resulting in a more specific estimation with respect to location. One possible limitation of the scaling procedure is that the network model is fitted on only the subset of regions. It does not take into account the regions that are outside the selected regions. This may lead to misleading results as we are considering an isolated network, whereas, in fact, the local network is part of the national network. This may be a topic for further research.

Considering our results, the question arises as to why the method does not work in certain circumstances and why it works very well in other cases. We believe that the underlying network, and whether that network captures the actually spreading mechanism of the outbreak, largely determines the performance of the method. One of the implicit assumptions of the network used in this article is that the outbreak spreads from a source and that the cases occur as the outbreak spreads. Over time more and more people are exposed as the food product is distributed across the country from the original source. This corresponds with the spreading mechanism of the *Salmonella* typhimurium outbreak, in which the contaminated food product was mainly sold in a single region, and as more and more people bought the product and transported it to their home region, the outbreak spread.

For the *Salmonella* thompson outbreak, the contaminated food product was not sold locally. Rather, the contamination source was a single factory, and only when the product had been distributed to supermarkets all over the country did the outbreak start. The first cases did not occur in the vicinity of the factory. Rather, the whole country was contaminated at once instead of the contamination spreading out from a single point. In reality, there were many (secondary) sources, namely all the supermarkets from which the cases bought the contaminated food product. The gravity model approximation in this article does not allow such one-time national contamination with many small sources.

Considering this situation, one might speculate that if the *Salmonella* thompson and *Escherichia coli* outbreaks were analyzed on European scale, we would be able to reconstruct

the origin of the outbreak. We, however, doubt that this would be the case. The spreading pattern of the two outbreaks would still be a simultaneous contamination over the whole of the Netherlands originating from multiple sources. This may contradict the spreading assumption underlying the gravity model approximation that assumes a single source from which the outbreak spreads. In this case, the needed approximation is very different, and the underlying network should capture only the human mobility pattern and not the food distribution network.

If real-world data were available, the gravity model could be fitted to obtain accurate parameter estimates. The origin of outbreaks such as *Salmonella* thompson and *Escherichia coli* could be reconstructed if detailed information of the food distribution network in the Netherlands were available, i.e., data on the distribution of food products from factories and packaging facilities to distribution centers and supermarkets.

We have seen that fitting the network model on different subsets of the data gives an indication of the performance of the method. This idea could be extended to develop a method that can be used to test whether the method works for a specific outbreak. The model can be fit to random subsets of the data to test the robustness of the model under these subsets. The extent to which the method estimates the outbreak origin consistently in the subsets may provide an indication of whether one can trust the estimation results. This is a topic for future research.

Future model improvements may consist of adapting the model structure and including real food distribution data. First, the network model can be adapted to allow for multilevel outbreaks with multiple (secondary) sources. At the top level, one may have a single primary source such as a factory or distribution center, which causes a national spreading pattern to multiple secondary sources. The secondary sources, such as supermarkets, constitute the second level of the network model. Each source in the second level can result in a local spreading pattern of the disease. Second, we believe that building the network model on real food distribution data instead of using the gravity model may contribute to substantial improvements in model performance.

Alternatively, one could consider other methods that do not only consider the shortest or highest probability paths along a network but all possible paths.[15]

CONCLUSIONS

We conclude that, although the method of reconstructing the outbreak origin developed by Manitz et al[6] has the potential to perform very well, care should be taken to not blindly use the gravity model as valid approximation for any food distribution pattern. The method can lead to misleading results and completely misconstruct the outbreak origin. Care should be taken as to whether the underlying network model correctly captures the spreading mechanism of the outbreak in terms of spatial scale and single or multiple source outbreak.

ACKNOWLEDGMENTS

We thank Emmanuel Lesaffre, Michael Höhle, and two anonymous reviewers for the suggestions and comments that improved this article.

REFERENCES

1. Hald T, Aspinall W, Devleesschauwer B, et al. World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. *PloS ONE*. 2016;110:e0145839.
2. Dwyer DM, Strickler H, Goodman RA, Armenian HK. Use of case-control studies in outbreak investigations. *Epidemiol Rev*. 1994;160:109–123.
3. Brandwagt D, van den Wijngaard C, Tulen AD, et al. Outbreak of *Salmonella* Bovismorbificans associated with the consumption of uncooked ham products, the Netherlands, 2016 to 2017. *Euro Surveill*. 2018;230:ISSN 15607917.
4. Friesema I, de Jong A, Hofhuis A, et al. Large outbreak of salmonella thompson related to smoked salmon in the Netherlands, August to December 2012. *Euro Surveill*. 2014;190ISSN 15607917.
5. Jacobs R, Lesaffre E, Teunis PF, Höhle M, van de Kassteele J. Identifying the source of food-borne disease outbreaks: an application of Bayesian variable selection. *Stat Methods Med Res*. 2019;280:1126–1140.
6. Manitz J, Kneib T, Schlather M, Helbing D,Brockmann D. Origin detection during food-borne disease outbreaks -A case study of the 2011 EHEC/HUS outbreak in Germany. *PLOS Currents*. 2014;6.
7. Anderson JE. A theoretical foundation for gravity equation. *American Economic Review*. 1979;690:106–116.
8. Weidlich H, Haag G, Weidlich W. A stochastic theory of interregional migration. *Geographical Analysis*. 1984;160:331–357.
9. Frank C, Werber D, Cramer JP, et al. Epidemic profile of Shiga-Toxin-Producing Escherichia coli O104:H4 outbreak in Germany. *N Engl J Med*. 2011;365: 1771–1780.
10. Van Duynhoven YT, Isken LD, Borgen K, et al. A prolonged outbreak of *Salmonella Typhimurium* infection related to an uncommon vehicle: hard cheese made from raw milk. *Epidemiol Infec*. 2009;137:1548–1557.
11. Friesema I, Sigmundsdottir G, van der Zwaluw K, et al. An international outbreak of Shi ga toxin-producing Escherichia coli O157 infection due to lettuce, September - October 2007. *Euro Surveill*. 2008;130:1–5.
12. Manitz J, Harbering J, Schmidt M, Kneib T, Schöbel A. Source estimation for propagation processes on complex networks with an application to delays in public transportation systems. *Journal of the Royal Statistical Society C*. 2017;660521–536.
13. Ángeles Serrano M, Boguñá M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad SciUSA*. 2009;1060:6483–6488.
14. Lawson AB. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. CRC Press, Boca Raton; 2009.
15. Horn AL, Friedrich H. Locating the source of large-scale outbreaks of foodborne disease. *J R Soc Interface*. 2019;160:20180624.